

# Discovering New Physics with Voronoi Tessellations

Dipsikha Debnath<sup>1,a)</sup>, James S. Gainer<sup>2</sup>, Doojin Kim<sup>1</sup> and Konstantin T. Matchev<sup>1</sup>

<sup>1</sup>*Physics Department, University of Florida, Gainesville, FL 32611, USA*

<sup>2</sup>*Dept. of Physics and Astronomy, University of Hawaii, Honolulu, HI 96822, USA*

<sup>a)</sup>dipsikha.debnath@gmail.com

**Abstract.** High energy experimental data can be viewed as a sampling of the relevant phase space. We point out that one can apply Voronoi tessellations in order to understand the underlying probability distributions in this phase space. Interesting features in the data can then be discovered by studying the properties of the ensemble of Voronoi cells. For illustration, we demonstrate the detection of kinematic “edges” in two dimensions, which may signal physics beyond the standard model. We motivate the algorithm with some analytical results derived for perfect lattices, and show that the method is further improved with the addition of a few Voronoi relaxation steps via Lloyd’s method.

## INTRODUCTION

In high energy physics, the data is a collection of “events”, which are distributed in phase space,  $\mathcal{P}$ , according to the differential cross-section

$$\frac{d\sigma}{d\vec{x}} \equiv f(\vec{x}, \{\alpha\}). \quad (1)$$

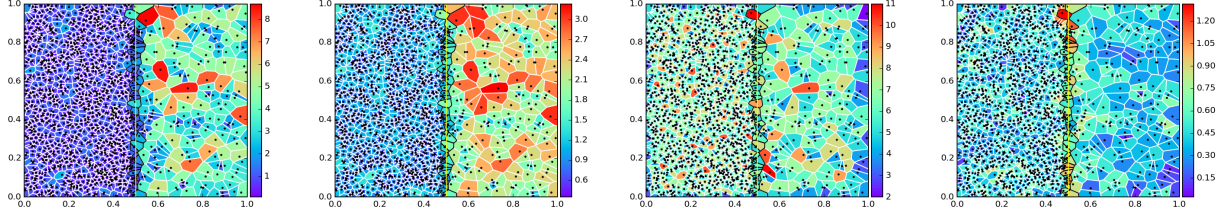
Here  $\vec{x} \in \mathcal{P}$  is a phase space point, which is often parameterized in terms of the momentum components of the final state particles. The set  $\{\alpha\}$  is a set of model parameters, e.g., particle masses, widths, couplings, etc. The function (1) consists of two contributions:

$$f(\vec{x}, \{\alpha\}) \equiv f_{SM}(\vec{x}, \{\alpha_{SM}\}) + f_{NP}(\vec{x}, \{\alpha_{NP}\}), \quad (2)$$

where  $f_{SM}$  represents the distribution expected from Standard Model (SM) processes, a.k.a. “the background”, while  $f_{NP}$  is the contribution due to new physics, i.e., “the signal”. A promising way to look for new physics is to identify structural features in the differential distributions of the observed events, which might be present in  $f_{NP}$ , but not in  $f_{SM}$ . This idea is similar to the bump-hunting technique in resonance searches, where we look for the Breit-Wigner peak in  $f_{NP}$  over the smooth background described by  $f_{SM}$ . Even when some of the decay products (e.g., neutrinos or dark matter particles) are invisible in the detector, one may still look for discontinuities or singularities [1] in the invariant mass distributions of the visible particles observed in the detector. Examples of such special features in  $f_{NP}$  include: kinematic endpoints [2, 3, 4, 5], kinematic boundaries [6, 7, 8, 9, 10], kinks [11, 12, 13, 14, 15] and cusps [16, 17, 18, 19]. These features are *not present* in the background distribution  $f_{SM}$ .

Here we concentrate on two-dimensional high energy particle physics data, but our study can be easily generalized to higher dimensions [20]. We assume that the signal distribution,  $f_{NP}$ , changes dramatically or has a discontinuity in phase space. Such a kinematic boundary or “edge” can reveal the existence of new particles. Edge detection has been studied in the experimental and observational sciences [21]. However, in particle physics, the standard methods of edge detection face several challenges, namely

1. *The data may be sparse.* Traditional edge detection methods focus on images, where each pixel contains a data point for a continuous variable (intensity). In contrast, in particle physics we look for an edge, which is a possible signature of new physics, with a comparatively small number of signal events.
2. *The analytic form of the distributions  $f_{SM}$  and  $f_{NP}$  describing the data may be unknown.* If the parametric form of the distribution (2) is known, we can promptly apply likelihood methods to determine edges. However, it is usually difficult to get an exact analytical form for  $f_{SM}$ , especially in the case of reducible backgrounds, where



**FIGURE 1.** Voronoi tessellations for 1400 data points selected from the probability density (3) with  $\rho = 6$ . The Voronoi polygons are color-coded by their area (left), perimeter (middle left), number of neighboring polygons (middle right) or scaled variance (4) (right).

detector effects play a major role. Moreover, we cannot be sure, *á priori*, that we have correctly assumed the specific new physics model [22]. Even if we have some idea of where the new physics edges may show up, a general procedure is always of greater practical value.

3. *The data may be in more than two dimensions.* As we mentioned above, edge detection is generally applied to two-dimensional images. However, in particle physics, multivariate analyses [23] are present everywhere. Therefore, in general we will be facing the problem of finding an  $(n - 1)$ -dimensional kinematic boundary in an  $n$ -dimensional parameter space.

Our proposed method for edge detection can handle all three of these challenges, and may become a useful tool for the experimental analyses in Run 2 of the CERN Large Hadron Collider (LHC).

## A Voronoi Method for Edge Detection

We start our analysis by making the Voronoi tessellation of some two-dimensional data, where each “event”,  $i$ , represents the corresponding generator point for the  $i^{\text{th}}$  Voronoi polygon [24, 25, 26]. This particular method of tessellation divides a given volume containing data points  $\{d_i\}$  into several regions,  $\mathcal{R}_i$ , such that each  $\mathcal{R}_i$  contains exactly one data point,  $d_i$ , and for any point  $p \in \mathcal{R}_i$ ,  $d_i$  is the nearest data point.

We focus to identify edge features such as discontinuities [27] without assuming the exact knowledge of the  $f_{NP}$  distribution. There exist several edge detection algorithms for binned data [28]. Our Voronoi method of edge detection avoids binning and includes the following steps:

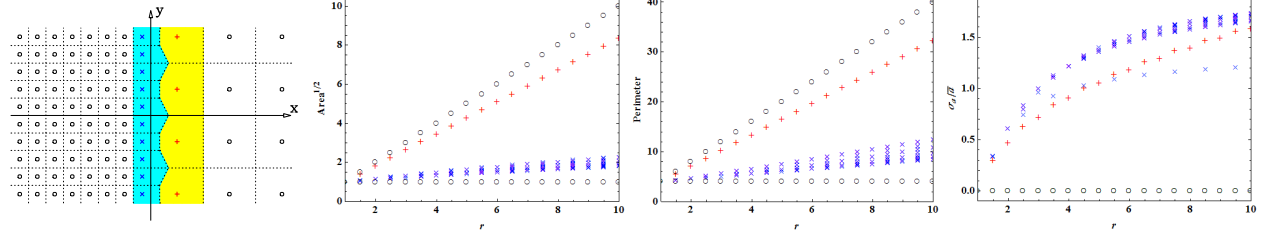
1. Construct the Voronoi tessellation for the data set.
2. Compute relevant attributes of the Voronoi cells.
3. (Optionally) use the information from the previous step to further process the data in some way.
4. Use some criterion to flag “candidate” edge cells.
5. Identify an edge from the collection of edge cell candidates.

Some useful intuition can be gained from the following toy example. We generate 1400 points according to the probability distribution

$$f(x, y) = \frac{2}{1 + \rho} [\rho H(0.5 - x) + H(x - 0.5)]. \quad (3)$$

within the unit square. In eq. (3),  $H(x)$  is the Heaviside step function and  $\rho$  is a constant density ratio. The resulting Voronoi tessellation is shown in Figure 1, where the color-code for each Voronoi polygon represents some standard property, such as area, perimeter, or number of immediate neighbors. The square is divided into left (L) and right (R) regions of constant, but unequal densities. Our goal is to spot the the vertical edge at  $x = 0.5$  (yellow solid line) where the density sharply changes from one region to other. For convenience, we outline the boundaries of the Voronoi cells, crossing the edge at  $x = 0.5$  as black and the remaining Voronoi cells away from the edge as white.

The two leftmost panels of Figure 1 show that the area and perimeter of the Voronoi polygons are somewhat correlated, while the middle right panel reveals that the typical number of nearest neighbors is similar in the two bulk



**FIGURE 2.** A regular lattice (5) generated for linear density ratio  $r = 3$  (left), and the dependence on  $r$  of several parameters of interest, namely cell area (middle left), cell perimeter (middle right) and scaled variance (right). Black circles indicate bulk cells, while blue  $\times$  (red  $+$ ) symbols denotes edge cells in the L (R) region.

regions. Therefore, these properties or aspects of Voronoi polygons cannot help in finding the edge cells (outlined in black). This is why we introduce a new variable, the scaled standard deviation of the areas of the neighboring cells,

$$\frac{\sigma_a}{\bar{a}} \equiv \frac{1}{\bar{a}} \sqrt{\sum_{n \in N_i} \frac{(a_n - \bar{a})^2}{|N_i| - 1}}, \quad (4)$$

where  $N_i$  is the set of neighbors of the  $i$ -th Voronoi polygon, and  $\bar{a}(N_i)$  is their mean area. The scaled standard deviation is quite successful in picking out edge cells and this can be visualized in the rightmost panel in Figure 1. Thus we choose (4) as our main selection variable<sup>1</sup>.

In order to understand the above results analytically, we consider a perfect grid of points which follows the probability distribution (3). The grid is generated by two integers  $n$  and  $m$  as

$$\vec{R} = [(n + 0.5) \hat{x} + (m + 0.5) \hat{y}] [H(-n) + rH(n)], \quad (5)$$

where the vectors  $\hat{x}$  and  $\hat{y}$  form an orthonormal basis and  $r \equiv \sqrt{\rho}$  is the corresponding linear density ratio. The left panel of Figure 2 shows an example grid for  $r = 3$ . We highlight the two columns of edge cells: in the L region (blue  $\times$  symbols) and the R region (red  $+$  symbols). The other three panels in Figure 2 show the behavior of some of their properties as a function of  $r$ . For the case of area and perimeter we notice that the values for edge cells are intermediate between the two bulk values. However, the scaled standard deviation is exactly zero for both bulk regions, and nonzero for the edge region, thus offering the possibility for good discrimination.

## Voronoi relaxation via Lloyd's algorithm

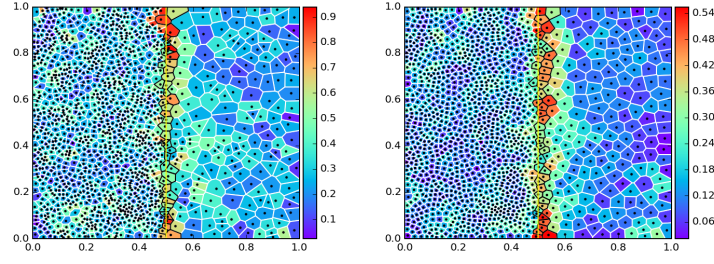
As we are dealing with a stochastic process, statistical fluctuations are unavoidable in the data. In particular, in Figure 1 we can easily spot a few bulk cells having relatively high values of  $\sigma_a/\bar{a}$ . This is why we introduce the idea of “smoothing” the data by applying a few iterations of Lloyd's algorithm [29], where at each iteration, the generator point is replaced by the centroid of the corresponding Voronoi cell.<sup>2</sup> Figure 3 shows the Voronoi tessellation after one (left panel) and five (right panel) Lloyd iterations. We find that the Voronoi polygons become more regularly shaped after relaxation and the fluctuations on each side of the boundary are washed out. Most importantly, the values of the scaled standard deviation (4) for the edge cells are enhanced relative to the rest.

Figure 3 also shows that as a result of the Voronoi relaxation, the data points from the dense L region flow towards the relatively sparse R region. Consequently, the edge cells with high  $\sigma_a/\bar{a}$  are displaced from their original locations (near the vertical yellow line). For this reason, once we select edge cell candidates after a certain number of Lloyd iterations, we need to trace them back to their original locations before doing any further quantitative data analysis.

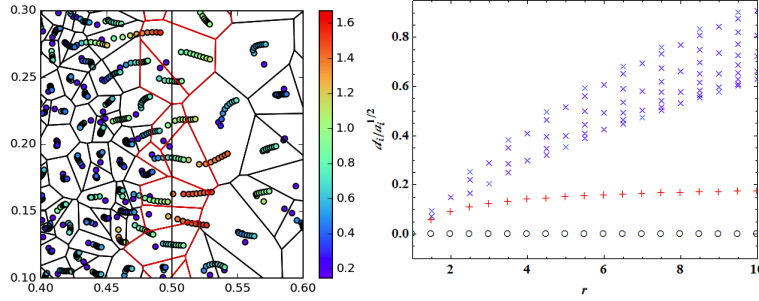
By comparing the displacements  $d_i$  of the generator points, we notice that the edge points tend to be displaced the farthest. We can use this as an alternative tagging method. To quantify this criterion, we define a dimensionless

<sup>1</sup>This is not the only option, however — we have investigated a number of other promising variables which will be discussed in a longer publication [20].

<sup>2</sup>An alternative approach, illustrated below in Figure 6, would be to leave the original Voronoi tessellation intact, but extend the calculation of (4) to include next-to-nearest neighbors, next-to-next-to-nearest neighbors, etc.



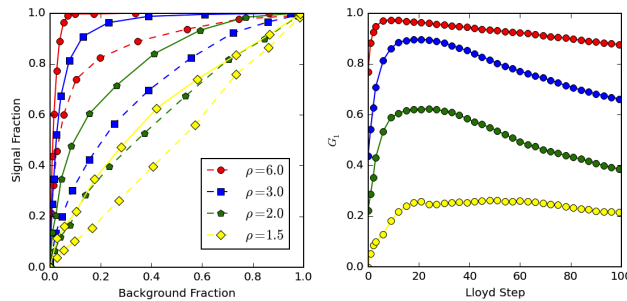
**FIGURE 3.** The evolution of the Voronoi tessellation shown in Figure 1 after one (left panel) and five (right panel) applications of Lloyd's algorithm. The cells are color-coded by the scaled variance (4).



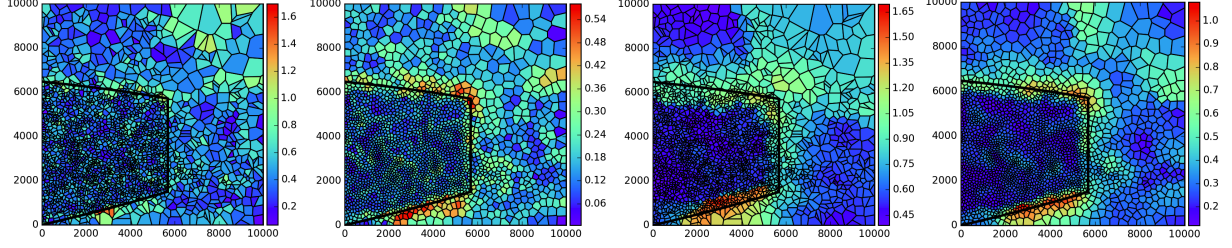
**FIGURE 4.** Left: a zoomed-in region near the vertical edge, which shows the originally generated points and their subsequent locations after repeated application of Lloyd's algorithm. The points are color-coded by scaled displacement,  $d_i / \sqrt{a_i}$ . Right: Predictions for  $d_i / \sqrt{a_i}$  after applying Lloyd's algorithm once, as a function of  $r$ , for the case of a regular lattice (5).

variable, the scaled displacement,  $d_i / \sqrt{a_i}$ , where we normalize by the square root of the cell area,  $a_i$ . The left panel in Figure 4 gives a closer view of one representative area near the edge and shows the result of several successive Lloyd iterations. The color code indicates that the scaled displacement is indeed a useful quantity, just like the scaled standard deviation (4). We confirm this by showing in the right panel of Figure 4 the exact result for the perfect grid (5).

We study the efficiency of our edge detection algorithm by analyzing ROC curves [30]. We generate a large dataset for (3), where we consider the edge cells as “signal” and the bulk cells as “background”. We plot the signal selection efficiency,  $\varepsilon_S$ , versus the background efficiency,  $\varepsilon_B$ , for different values of the minimum cut on the variable (4). Several  $\varepsilon_S(\varepsilon_B)$  curves, for different values of the density ratio  $\rho$ , and either with (solid) or without (dashed) Lloyd relaxation are shown in Figure 5. The ROC curves reveal that the algorithm is more efficient for higher density contrasts between the two regions. In addition, the Voronoi relaxation leads to a significant improvement of the result.



**FIGURE 5.** Left: ROC curves  $\varepsilon_S(\varepsilon_B)$  obtained using (4) as the discriminating variable. Right: The Gini index (6) found from the ROC curve obtained after the given number of Lloyd iterations.



**FIGURE 6.** Voronoi tessellations for the supersymmetry example described in the text.

The accuracy of our selection criteria is quantified by using the standard area under the curve [31] (AUROC) as represented by the Gini coefficient

$$G_1 \equiv 2 \text{AUROC} - 1 = 2 \int_0^1 d\varepsilon_B \times \varepsilon_S(\varepsilon_B) - 1, \quad (6)$$

where a value of 1 is obtained from the ROC curve of a perfectly discriminating variable, while a value of 0 corresponds to a totally random selection of events. The right panel of Figure 5 shows the dependence of  $G_1$  on the number of Lloyd steps. We see that the sensitivity improves dramatically within the first few iterations, and reaches an optimum plateau, after which the power of the test is degraded as the Voronoi grid begins to asymptote to the centroidal tessellation.

### An example from supersymmetry

We apply our proposed edge detection method to a standard benchmark example from supersymmetry; squark pair production at the 13 TeV LHC. We consider events where one squark undergoes a long cascade decay through a heavy neutralino,  $\tilde{\chi}_2^0$ , a slepton,  $\tilde{\ell}$ , and a light neutralino,  $\tilde{\chi}_1^0$ ; while the other decays directly to  $\tilde{\chi}_1^0$ . The mass spectrum is chosen to be  $m_{\tilde{q}} = 400$  GeV,  $m_{\tilde{\chi}_2^0} = 300$  GeV,  $m_{\tilde{\ell}} = 280$  GeV, and  $m_{\tilde{\chi}_1^0} = 200$  GeV. The invariant mass distributions of the final state particles, the two jets and the two leptons, exhibit kinematic edges.

In particular, here we consider the dilepton invariant mass,  $m_{\ell\ell}$ , and the three-body jet-lepton-lepton invariant mass,  $m_{j\ell\ell}$ . In Figure 6 we use the  $(m_{\ell\ell}^2, (m_{j\ell\ell}^2 - m_{\ell\ell}^2)/6)$  plane for plotting convenience. The solid black line in Figure 6 [8, 32] marks the location of the kinematic endpoint for signal events with the correct jet assignment. (The lack of knowledge of the charge of the jet creates a two-fold combinatorial ambiguity. Thus, for each event there are two entries in the plot.) The main SM background from  $t\bar{t}$  dilepton events is also included here.

In Figure 6, the Voronoi cells are color coded by their scaled standard deviation (4). In the left panel we exhibit the original data, while in the middle left panel we show the data after 5 Lloyd iterations. We reconsider the original data and extend the calculation of (4) including up to 5 tiers of nearest neighbors, showing the resulting plot in the middle right panel. We observe that either Voronoi relaxation or the addition of more tiers of neighboring cells reduces the fluctuation and sharpens the edge. Finally, in the rightmost panel of Figure 6 we show the result after 3 Lloyd iterations *and* also including 3 tiers of neighbors in the calculation of (4).

### Summary

We argue that the discovery of new kinematic features is an essential step in the discovery of physics beyond the standard model at the LHC and advocate the use of Voronoi methods for this purpose. The great flexibility of Voronoi methods is a blessing for the experimentalist; many useful properties of the Voronoi cells can be used to construct powerful variables tailored to specific new physics scenarios. A voluminous, quantitative study of the many options available to the experimenter will be presented in a companion paper [20].

## Acknowledgements

We thank S. Das, C. Kilic, Z. Liu, R. Lu, P. Ramond, X. Tata, J. Thaler, B. Tweedie, and D. Yaylali for useful discussions. Work supported in part by U.S. Department of Energy, in part by Grant DE-SC0010296. DK acknowledges support by LHC-TI postdoctoral fellowship under grant NSF-PHY-0969510.

## REFERENCES

- [1] I. W. Kim, Phys. Rev. Lett. **104**, 081601 (2010) [arXiv:0910.1149 [hep-ph]].
- [2] I. Hinchliffe, F. E. Paige, M. D. Shapiro, J. Soderqvist and W. Yao, Phys. Rev. D **55**, 5520 (1997) [hep-ph/9610544].
- [3] W. S. Cho, J. E. Kim and J. H. Kim, Phys. Rev. D **81**, 095010 (2010) [arXiv:0912.2354 [hep-ph]].
- [4] A. J. Barr and C. G. Lester, J. Phys. G **37**, 123001 (2010) [arXiv:1004.2732 [hep-ph]].
- [5] A. J. Barr, T. J. Khoo, P. Konar, K. Kong, C. G. Lester, K. T. Matchev and M. Park, Phys. Rev. D **84**, 095031 (2011) [arXiv:1105.2977 [hep-ph]].
- [6] D. Costanzo and D. R. Tovey, JHEP **0904**, 084 (2009) [arXiv:0902.2331 [hep-ph]].
- [7] M. Burns, K. T. Matchev and M. Park, JHEP **0905**, 094 (2009) [arXiv:0903.4371 [hep-ph]].
- [8] K. T. Matchev, F. Moortgat, L. Pape and M. Park, JHEP **0908**, 104 (2009) [arXiv:0906.2417 [hep-ph]].
- [9] K. T. Matchev and M. Park, Phys. Rev. Lett. **107**, 061801 (2011) [arXiv:0910.1584 [hep-ph]].
- [10] P. Agrawal, C. Kilic, C. White and J. H. Yu, Phys. Rev. D **89**, no. 1, 015021 (2014) [arXiv:1308.6560 [hep-ph]].
- [11] W. S. Cho, K. Choi, Y. G. Kim and C. B. Park, Phys. Rev. Lett. **100**, 171801 (2008) [arXiv:0709.0288 [hep-ph]].
- [12] B. Gripaios, JHEP **0802**, 053 (2008) [arXiv:0709.2740 [hep-ph]].
- [13] A. J. Barr, B. Gripaios and C. G. Lester, JHEP **0802**, 014 (2008) [arXiv:0711.4008 [hep-ph]].
- [14] W. S. Cho, K. Choi, Y. G. Kim and C. B. Park, JHEP **0802**, 035 (2008) [arXiv:0711.4526 [hep-ph]].
- [15] M. Burns, K. Kong, K. T. Matchev and M. Park, JHEP **0903**, 143 (2009) [arXiv:0810.5576 [hep-ph]].
- [16] T. Han, I. W. Kim and J. Song, Phys. Lett. B **693**, 575 (2010) [arXiv:0906.5009 [hep-ph]].
- [17] K. Agashe, D. Kim, M. Toharia and D. G. E. Walker, Phys. Rev. D **82**, 015007 (2010) [arXiv:1003.0899 [hep-ph]].
- [18] T. Han, I. W. Kim and J. Song, Phys. Rev. D **87**, no. 3, 035003 (2013) [arXiv:1206.5633 [hep-ph]].
- [19] T. Han, I. W. Kim and J. Song, Phys. Rev. D **87**, no. 3, 035004 (2013) [arXiv:1206.5641 [hep-ph]].
- [20] D. Debnath, J. S. Gainer, D. Kim and K. T. Matchev, (work in progress).
- [21] See, e.g., Davies, E. R., “Computer & Machine Vision: Theory, Algorithms, Practicalities”, Academic Press; 4 edition (March 19, 2012).
- [22] D. Debnath, J. S. Gainer and K. T. Matchev, Phys. Lett. B **743**, 1 (2015) [arXiv:1405.5879 [hep-ph]].
- [23] P. C. Bhat, Ann. Rev. Nucl. Part. Sci. **61**, 281 (2011).
- [24] G. Voronoi, Journal für die Reine und Angewandte Mathematik, **133**, 97 (1908).
- [25] G. L. Dirichlet, Journal für die Reine und Angewandte Mathematik, **40**, 209 (1850).
- [26] See, e.g., S. Okabe, B. Boots and K. Sugihara, “Spatial Tessellations: Concepts and Applications of Voronoi Diagrams,” John Wiley & Sons, 1992.
- [27] For a study in one dimension, see D. Curtin, Phys. Rev. D **85**, 075004 (2012) [arXiv:1112.1095 [hep-ph]].
- [28] See, e.g., J. Canny, IEEE Trans. Pattern Analysis and Machine Intelligence, 8(6):679-698, (1986).
- [29] S. P. Lloyd, IEEE Trans. on Information Theory, 28 (2): 129-137, (1982).
- [30] T. Fawcett, “An introduction to ROC analysis,”
- [31] J. Hanley and B. McNeil, Radiology, **143** (1), 29-36 (1982).
- [32] C. G. Lester, M. A. Parker and M. J. White, JHEP **0710**, 051 (2007) [hep-ph/0609298].